

A Constant-Factor Bi-Criteria Approximation Guarantee for k -means++

Dennis Wei

May 18, 2016

Abstract

This paper studies the k -means++ algorithm for clustering as well as the class of D^ℓ sampling algorithms to which k -means++ belongs. It is shown that for any constant factor $\beta > 1$, selecting βk cluster centers by D^ℓ sampling yields a constant-factor approximation to the optimal clustering with k centers, in expectation and without conditions on the dataset. This result extends the previously known $O(\log k)$ guarantee for the case $\beta = 1$ to the constant-factor bi-criteria regime. It also improves upon an existing constant-factor bi-criteria result that holds only with constant probability.

1 Introduction

The k -means problem and its variants constitute one of the most popular paradigms for clustering [Jain, 2010]. Given a set of n data points, the task is to group them into k clusters, each defined by a cluster center, such that the sum of distances from points to cluster centers (raised to a power ℓ) is minimized. Optimal clustering in this sense is known to be NP-hard, in particular for k -means ($\ell = 2$) [Dasgupta, 2008, Aloise et al., 2009, Mahajan et al., 2009, Awasthi et al., 2015] and k -medians ($\ell = 1$) [Jain et al., 2002]. In practice, the most widely used algorithm remains Lloyd’s [1957, 1982] (often referred to as the k -means algorithm), which alternates between updating centers given cluster assignments and re-assigning points to clusters.

In this paper, we study an enhancement to Lloyd’s algorithm known as k -means++ [Arthur and Vassilvitskii, 2007] and the more general class of D^ℓ sampling algorithms to which k -means++ belongs. These algorithms select cluster centers randomly from the given data points with probabilities proportional to their current costs. The clustering can then be refined using Lloyd’s algorithm. D^ℓ sampling is attractive for two reasons: First, it is guaranteed to yield an expected $O(\log k)$ approximation to the optimal clustering with k centers [Arthur and Vassilvitskii, 2007]. Second, it is as simple as Lloyd’s algorithm, both conceptually as well as computationally with $O(nkd)$ running time in d dimensions.

The particular focus of this paper is on the setting where an optimal k -clustering remains the benchmark but more than k cluster centers can be sampled to improve the approximation. Specifically, it is shown (see Theorem 1 and Corollary 1) that for any constant factor $\beta > 1$, if βk centers are chosen by D^ℓ sampling, then a constant-factor approximation to the optimal k -clustering is obtained. This guarantee holds in expectation and for all datasets, like the one in Arthur and Vassilvitskii [2007], and improves upon the $O(\log k)$ factor therein. Such a result is known as a constant-factor bi-criteria approximation since both the optimal cost and the relevant degrees of freedom (k in this case) are exceeded but only by constant factors.

In the context of clustering, bi-criteria approximations can be valuable because an appropriate number of clusters k is almost never known or pre-specified in practice. Approaches to determining k from the data are all ideally based on knowing how the optimal cost decreases as k increases, but obtaining this optimal trade-off between cost and k is NP-hard as mentioned earlier. Alternatively, a simpler algorithm that has a constant-factor bi-criteria guarantee would ensure that the trade-off curve generated by this algorithm

deviates by no more than constant factors along both axes from the optimal curve. This may be more appealing than a deviation along the cost axis that grows with k . Furthermore, if a solution with a specified number of clusters k is truly required, then linear programming techniques can be used to select a k -subset from the βk cluster centers while still maintaining a constant-factor approximation [Aggarwal et al., 2009, Charikar et al., 2002].

The main result in this paper differs from the constant-factor bi-criteria approximation established in Aggarwal et al. [2009] in that the latter holds only with constant probability as opposed to in expectation. Using Markov’s inequality, a constant-probability corollary can be derived from Theorem 1 herein, and doing so improves upon the approximation factor of Aggarwal et al. [2009] by more than a factor of 2. The present paper also differs from recent work on more general bi-criteria approximation of k -means by Makarychev et al. [2015], which analyzes substantially more complex algorithms.

In the next section, existing work on D^ℓ sampling and clustering approximations in general is reviewed in more detail. Section 2 gives a formal statement of the problem, the D^ℓ sampling algorithm, and existing lemmas regarding the algorithm. Section 3 states the main results of the paper and compares them to previous results. Proofs are presented in Section 4 and the paper concludes in Section 5.

1.1 Related Work

There is a considerable literature on approximation algorithms for k -means, k -medians, and related problems, spanning a wide range in the trade-off between tighter approximation factors and lower algorithm complexity. At one end, exact algorithms [Inaba et al., 1994] and several polynomial-time approximation schemes (PTAS) [Matoušek, 2000, Badoiu et al., 2002, de la Vega et al., 2003, Har-Peled and Mazumdar, 2004, Kumar et al., 2010, Chen, 2009, Feldman et al., 2007, Jaiswal et al., 2014] have been proposed for k -means and k -medians. While these have polynomial running times in n , the dependence on k and sometimes on the dimension d is exponential or worse. A simpler local search algorithm was shown to yield a $((3 + 2/p)^\ell + \epsilon)$ approximation for k -means ($\ell = 2$) in Kanungo et al. [2004] and k -medians ($\ell = 1$) in Arya et al. [2004], the latter under the additional constraint that centers are chosen from a finite set. This local search however requires a polynomial number of iterations of complexity $n^{O(p)}$, and Kanungo et al. [2004] also rely on a discretization to an ϵ -approximate centroid set [Matoušek, 2000] of size $O(n\epsilon^{-d} \log(1/\epsilon))$. Linear programming algorithms offer similar constant-factor guarantees with similar running times for k -medians (again the finite set variant) and the related problem of facility location [Charikar et al., 2002, Jain and Vazirani, 2001].

In contrast to the above, this paper focuses on simpler algorithms in the D^ℓ sampling class, including k -means++. In Arthur and Vassilvitskii [2007], it was proved that D^ℓ sampling results in an $O(\log k)$ approximation, in expectation and for all datasets. The current work builds upon Arthur and Vassilvitskii [2007] to extend the guarantee to the constant-factor bi-criteria regime. Arthur and Vassilvitskii [2007] also provided a matching lower bound, exhibiting a dataset on which k -means++ achieves an expected $\Omega(\log k)$ approximation.

Sampling algorithms have been shown to yield improved $O(1)$ approximation factors provided that the dataset satisfies certain conditions. Such a result was established in Ostrovsky et al. [2012] for k -means++ and other variants of Lloyd’s algorithm under the condition that the dataset is well-suited in a sense to partitioning into k clusters. In Mettu and Plaxton [2004], an $O(1)$ approximation was shown for a somewhat more complicated algorithm called successive sampling with $O(n(k + \log n) + k^2 \log^2 n)$ running time, subject to a bound on the dispersion of the points. A constant-factor approximation with slightly superlinear running time has also been obtained in the streaming setting [Guha et al., 2003].

For k -means++, the $\Omega(\log k)$ lower bound in Arthur and Vassilvitskii [2007], which holds in expectation, has spurred follow-on works on the question of whether k -means++ might guarantee a constant-factor approximation with reasonably large probability. Negative answers were provided by Brunsch and Röglin

[2013], who showed that an approximation factor better than $(2/3) \log k$ cannot be achieved with probability higher than a decaying exponential in k , and Bhattacharya et al. [2014], who showed that a similar statement holds even in 2 dimensions.

In a similar direction to the one pursued in the present work, Aggarwal et al. [2009] showed that if the number of cluster centers can be increased to a constant factor times k , then a constant-factor approximation can be achieved with constant probability. Specifically, they prove that using $\lceil 16(k + \sqrt{k}) \rceil$ centers gives an approximation factor of 20 with probability 0.03, together with a general bi-criteria guarantee but without explicit constants. An $O(1)$ factor was also obtained independently by Ailon et al. [2009] using more centers, of order $O(k \log k)$. As mentioned, the result of Aggarwal et al. [2009] differs from Theorem 1 herein in being true with constant probability as opposed to in expectation. Furthermore, Section 3.1 shows that a constant-probability corollary of Theorem 1 improves significantly upon Aggarwal et al. [2009].

Recently, Makarychev et al. [2015] has also established constant-factor bi-criteria results for k -means. Their work differs from the present paper in studying more complex algorithms. First, similar to Kanungo et al. [2004], Makarychev et al. [2015] reduce the k -means problem to an ϵ -approximate, finite-set instance of k -medians of size $n^{O(\log(1/\epsilon)/\epsilon^2)}$. Subsequently, linear programming and local search algorithms are considered, the latter the same as in Kanungo et al. [2004], Arya et al. [2004], and both with polynomial complexity in the size of the k -medians instance.

2 Preliminaries

2.1 Problem Definition

We are given n points x_1, \dots, x_n in a real metric space \mathcal{X} with metric $D(x, y)$. The objective is to choose t cluster centers c_1, \dots, c_t in \mathcal{X} and assign points to the nearest cluster center to minimize the potential function

$$\phi = \sum_{i=1}^n \min_{j=1, \dots, t} D(x_i, c_j)^\ell. \quad (1)$$

A cluster is thus defined by the points x_i assigned to a center c_j , where ties (multiple closest centers) are broken arbitrarily. For a subset of points \mathcal{S} , define $\phi(\mathcal{S}) = \sum_{x_i \in \mathcal{S}} \min_{j=1, \dots, t} D(x_i, c_j)^\ell$ to be the contribution to the potential from \mathcal{S} ; $\phi(x_i)$ is the contribution from a single point x_i .

The exponent $\ell \geq 1$ in (1) is regarded as a problem parameter. Letting $\ell = 2$ and D be Euclidean distance, we have what is usually known as the k -means problem, so-called because the optimal cluster centers are means of the points assigned to them. The choice $\ell = 1$ is also popular and corresponds to the k -medians problem.

Throughout this paper, an optimal clustering will always refer to one that minimizes (1) over solutions with $t = k$ clusters, where $k \geq 2$ is given. Likewise, the term optimal cluster and symbol \mathcal{A} will refer to one of the k clusters from this optimal solution. The goal is to approximate the potential ϕ^* of this optimal k -clustering using $t = \beta k$ cluster centers for $\beta \geq 1$.

2.2 D^ℓ Sampling Algorithm

The D^ℓ sampling algorithm chooses cluster centers randomly from x_1, \dots, x_n with probabilities proportional to their current contributions to the potential, as detailed in Algorithm 1. Following Arthur and Vassilvitskii [2007], the case $\ell = 2$ is referred to as the k -means++ algorithm and the probabilities used after the first iteration are referred to as D^2 weighting (hence D^ℓ in general). For t cluster centers, the running time of D^ℓ sampling is $O(ntd)$ in d dimensions.

Algorithm 1 D^ℓ Sampling

Input: Data points x_1, \dots, x_n , number of clusters t
Select first cluster center c_1 uniformly at random from x_1, \dots, x_n .
Compute $\phi(x_i)$ for $i = 1, \dots, n$.
for $j = 2$ **to** t **do**
 Select j th center $c_j = x_i$ with probability $\phi(x_i)/\phi$.
 Update $\phi(x_i)$ for $i = 1, \dots, n$.

In practice, Algorithm 1 is used as an initialization to Lloyd's algorithm, which usually produces further decreases in the potential. The analysis herein pertains only to Algorithm 1 and not to the subsequent improvement due to Lloyd's algorithm.

2.3 Existing Lemmas Regarding D^ℓ Sampling

The following lemmas synthesize results from Arthur and Vassilvitskii [2007] that bound the expected potential within a single optimal cluster due to selecting a center from that cluster with uniform or D^ℓ weighting, as in Algorithm 1. These lemmas define the constant $r_D^{(\ell)}$ appearing in the main results below and are also used in their proof.

Lemma 1. [Arthur and Vassilvitskii, 2007, Lemmas 3.1 and 5.1] *Given an optimal cluster \mathcal{A} , let ϕ be the potential resulting from selecting a first cluster center randomly from \mathcal{A} with uniform weighting. Then $\mathbb{E}[\phi(\mathcal{A})] \leq r_u^{(\ell)} \phi^*(\mathcal{A})$ for any \mathcal{A} , where*

$$r_u^{(\ell)} = \begin{cases} 2, & \ell = 2 \text{ and } D \text{ is Euclidean,} \\ 2^\ell, & \text{otherwise.} \end{cases}$$

Lemma 2. [Arthur and Vassilvitskii, 2007, Lemma 3.2] *Given an optimal cluster \mathcal{A} and an initial potential ϕ , let ϕ' be the potential resulting from adding a cluster center selected randomly from \mathcal{A} with D^ℓ weighting. Then $\mathbb{E}[\phi'(\mathcal{A})] \leq r_D^{(\ell)} \phi^*(\mathcal{A})$ for any \mathcal{A} , where $r_D^{(\ell)} = 2^\ell r_u^{(\ell)}$.*

The factor of 2^ℓ between $r_u^{(\ell)}$ and $r_D^{(\ell)}$ for general ℓ is explained just before Theorem 5.1 in Arthur and Vassilvitskii [2007].

3 Main Results

The main results of this paper are stated below in terms of the single-cluster approximation ratio $r_D^{(\ell)}$ defined by Lemma 2. Subsequently in Section 3.1, the results are discussed in the context of previous work.

Theorem 1. *Let ϕ be the potential resulting from selecting βk cluster centers according to Algorithm 1, where $\beta \geq 1$. The expected approximation ratio is then bounded as*

$$\frac{\mathbb{E}[\phi]}{\phi^*} \leq r_D^{(\ell)} \left(1 + \min \left\{ \frac{\varphi(k-2)}{(\beta-1)k + \varphi}, H_{k-1} \right\} \right) - \Theta \left(\frac{1}{n} \right),$$

where $\varphi = (1 + \sqrt{5})/2 \doteq 1.618$ is the golden ratio and $H_k = 1 + \frac{1}{2} + \dots + \frac{1}{k} \sim \log k$ is the k th harmonic number.

In the proof of Theorem 1 in Section 4.2, it is shown that the $1/n$ term is indeed non-positive and can therefore be omitted, with negligible loss for large n .

The approximation ratio bound in Theorem 1 is stated as a function of k . The following corollary confirms that the theorem also implies a constant-factor bi-criteria approximation.

Corollary 1. *With the same definitions as in Theorem 1, the expected approximation ratio is bounded as*

$$\frac{\mathbb{E}[\phi]}{\phi^*} \leq r_D^{(\ell)} \left(1 + \frac{\varphi}{\beta - 1} \right).$$

Proof. The minimum appearing in Theorem 1 is bounded from above by its first term. This term is in turn increasing in k with asymptote $\varphi/(\beta - 1)$, which can therefore be taken as a k -independent bound. \square

It follows from Corollary 1 that a constant “oversampling” ratio $\beta > 1$ leads to a constant-factor approximation. Theorem 1 offers a further refinement for finite k .

The bounds in Theorem 1 and Corollary 1 consist of two factors. As β increases, the second, parenthesized factor decreases to 1 either exactly or approximately as $1/(\beta - 1)$. The first factor of $r_D^{(\ell)}$ however is no smaller than 4, and is a direct consequence of Lemma 2. Any improvement of Lemma 2 would therefore strengthen the approximation factors above. This subject is briefly discussed in Section 5.

3.1 Comparisons to Existing Results

A comparison of Theorem 1 to results in Arthur and Vassilvitskii [2007] is implicit in its statement since the H_{k-1} term in the minimum comes directly from Arthur and Vassilvitskii [2007, Theorems 3.1 and 5.1]. For $k = 2, 3$, the first term in the minimum is smaller than H_{k-1} for any $\beta \geq 1$, and hence Theorem 1 is always an improvement. For $k > 3$, Theorem 1 improves upon Arthur and Vassilvitskii [2007] for β greater than the critical value

$$\beta_c = 1 + \frac{\phi(k - 2 - H_{k-1})}{kH_{k-1}}.$$

Numerical evaluation of β_c shows that it reaches a maximum value of 1.204 at $k = 22$ and then decreases back toward 1 roughly as $1/H_{k-1}$. It can be concluded that for any k , at most 20% oversampling is required for Theorem 1 to guarantee a better approximation than Arthur and Vassilvitskii [2007].

The most closely related result to Theorem 1 and Corollary 1 is found in Aggarwal et al. [2009, Theorem 1]. The latter establishes a constant-factor bi-criteria approximation that holds with constant probability, as opposed to in expectation. Since a bound on the expectation implies a bound with constant probability via Markov’s inequality, a direct comparison with Aggarwal et al. [2009] is possible. Specifically, for $\ell = 2$ and the $t = \lceil 16(k + \sqrt{k}) \rceil$ cluster centers assumed in Aggarwal et al. [2009], Theorem 1 in the present work implies that

$$\begin{aligned} \frac{\mathbb{E}[\phi]}{\phi^*} &\leq 8 \left(1 + \min \left\{ \frac{\varphi(k - 2)}{\lceil 15k + 16\sqrt{k} \rceil + \varphi}, H_{k-1} \right\} \right) \\ &\leq 8 \left(1 + \frac{\varphi}{15} \right), \end{aligned}$$

after taking $k \rightarrow \infty$. Then by Markov’s inequality,

$$\frac{\phi}{\phi^*} \leq \frac{8}{0.97} \left(1 + \frac{\varphi}{15} \right) \doteq 9.137$$

with probability at least $1 - 0.97 = 0.03$ as in Aggarwal et al. [2009]. This 9.137 approximation factor is less than half the factor of 20 in Aggarwal et al. [2009].

Corollary 1 may also be compared to the results in Makarychev et al. [2015], although it should be re-emphasized that the latter analyzes different, substantially more complex algorithms, with running time at least $n^{O(\log(1/\epsilon)/\epsilon^2)}$ for reasonably small ϵ . The main difference between Corollary 1 and the bounds in Makarychev et al. [2015] is the extra factor of $r_D^{(\ell)}$ since the factor of $1 + \phi/(\beta - 1)$ is comparable, at least for moderate values of β that are of practical interest. As discussed above and in Section 5, the factor of $r_D^{(\ell)}$ is due to Lemma 2 and is unlikely to be intrinsic to the D^ℓ sampling algorithm.

4 Proofs

The overall strategy used to prove Theorem 1 is similar to that in Arthur and Vassilvitskii [2007]. The key intermediate result is Lemma 3 below, which relates the potential at a later iteration in Algorithm 1 to the potential at an earlier iteration. Section 4.1 is devoted to proving Lemma 3. Subsequently in Section 4.2, Theorem 1 is proven by an application of Lemma 3.

In the sequel, we say that an optimal cluster \mathcal{A} is covered by a set of cluster centers if at least one of the centers lies in \mathcal{A} . Otherwise \mathcal{A} is uncovered. Also define $\rho = r_D^{(\ell)}\phi^*$ as an abbreviation.

Lemma 3. *For an initial set of centers leaving u optimal clusters uncovered, let ϕ denote the potential, \mathcal{U} the union of uncovered clusters, and \mathcal{V} the union of covered clusters. Let ϕ' denote the potential resulting from adding $t \geq u$ centers, each selected randomly with D^ℓ weighting as in Algorithm 1. Then the new potential is bounded in expectation as*

$$\mathbb{E}[\phi' \mid \phi] \leq c_{\mathcal{V}}(t, u)\phi(\mathcal{V}) + c_{\mathcal{U}}(t, u)\rho(\mathcal{U})$$

for coefficients $c_{\mathcal{V}}(t, u)$ and $c_{\mathcal{U}}(t, u)$ that depend only on t, u . This holds in particular for

$$c_{\mathcal{V}}(t, u) = 1 + \frac{\varphi u}{t - u + \varphi}, \tag{2a}$$

$$c_{\mathcal{U}}(t, u) = \begin{cases} 1 + \frac{\varphi(u-1)}{t - u + \varphi}, & u > 0, \\ 0, & u = 0. \end{cases} \tag{2b}$$

4.1 Proof of Lemma 3

Lemma 3 is proven using induction, showing that if it holds for (t, u) and $(t, u + 1)$, then it also holds for $(t + 1, u + 1)$, similar to the proof of Arthur and Vassilvitskii [2007, Lemma 3.3]. The proof is organized into three parts. Section 4.1.1 provides base cases. In Section 4.1.2, sufficient conditions on the coefficients $c_{\mathcal{V}}(t, u)$, $c_{\mathcal{U}}(t, u)$ are derived that allow the inductive step to be completed. In Section 4.1.3, it is shown that the closed-form expressions in (2) are consistent with the base cases in Section 4.1.1 and satisfy the sufficient conditions from Section 4.1.2, thus completing the proof.

4.1.1 Base cases

This subsection exhibits two base cases of Lemma 3. While the second of these base cases does not conform to the functional forms in (2), it is shown later in Section 4.1.3 that the same base cases also hold with coefficients given by (2).

The first case corresponds to $u = 0$, for which we have $\phi(\mathcal{V}) = \phi$. Since adding centers cannot increase the potential, i.e. $\phi' \leq \phi$ deterministically, Lemma 3 holds with

$$c_{\mathcal{V}}(t, 0) = 1, \quad c_{\mathcal{U}}(t, 0) = 0, \quad t \geq 0. \tag{3}$$

The second base case occurs for $t = u$, $u \geq 1$. For this purpose, a slightly strengthened version of Arthur and Vassilvitskii [2007, Lemma 3.3] is used, as given next.

Lemma 4. *With the same definitions as in Lemma 3 except with $t \leq u$, we have*

$$\mathbb{E}[\phi' \mid \phi] \leq (1 + H_t)\phi(\mathcal{V}) + (1 + H_{t-1})\rho(\mathcal{U}) + \frac{u-t}{u}\phi(\mathcal{U}),$$

where we define $H_0 = 0$ and $H_{-1} = -1$ for convenience.

The improvement is in the coefficient in front of $\rho(\mathcal{U})$, from $(1 + H_t)$ to $(1 + H_{t-1})$. The proof follows that of Arthur and Vassilvitskii [2007, Lemma 3.3] with some differences and is deferred to Appendix A.

Specializing to the case $t = u$, Lemma 4 coincides with Lemma 3 with coefficients

$$c_{\mathcal{V}}(u, u) = 1 + H_u, \quad c_{\mathcal{U}}(u, u) = 1 + H_{u-1}. \quad (4)$$

4.1.2 Sufficient conditions on coefficients

In this subsection, it is assumed inductively that Lemma 3 holds for (t, u) and $(t, u + 1)$. The induction to the case $(t + 1, u + 1)$ is then completed under the following sufficient conditions on the coefficients:

$$c_{\mathcal{V}}(t, u + 1) \geq 1, \quad (5a)$$

$$(c_{\mathcal{V}}(t, u + 1) - c_{\mathcal{U}}(t, u + 1))c_{\mathcal{V}}(t, u)^2 \geq (c_{\mathcal{U}}(t, u + 1) - c_{\mathcal{V}}(t, u))^2, \quad (5b)$$

and

$$c_{\mathcal{V}}(t + 1, u + 1) \geq \frac{1}{2} \left[c_{\mathcal{V}}(t, u) + (c_{\mathcal{V}}(t, u))^2 + 4 \max\{c_{\mathcal{V}}(t, u + 1) - c_{\mathcal{V}}(t, u), 0\}^{1/2} \right], \quad (6a)$$

$$c_{\mathcal{U}}(t + 1, u + 1) \geq c_{\mathcal{V}}(t, u). \quad (6b)$$

The first pair of conditions (5) applies to the coefficients involved in the inductive hypothesis for (t, u) and $(t, u + 1)$. The second pair (6) can be seen as a recursive specification of the new coefficients for $(t + 1, u + 1)$. This inductive step together with base cases (3) and (4) are sufficient to extend Lemma 3 to all $t > u$, starting with $(t + 1, u + 1) = (2, 1)$ from $(t, u) = (1, 0)$ and $(t, u + 1) = (1, 1)$.

The inductive step is broken down into a series of three lemmas, each building upon the last. The first lemma applies the inductive hypothesis to derive a bound on the potential that depends not only on $\phi(\mathcal{V})$ and $\rho(\mathcal{U})$ but also on $\phi(\mathcal{U})$.

Lemma 5. *Assume that Lemma 3 holds for (t, u) and $(t, u + 1)$. Then for the case $(t + 1, u + 1)$, i.e. ϕ corresponding to $u + 1$ uncovered clusters and ϕ' resulting after adding $t + 1$ centers,*

$$\mathbb{E}[\phi' \mid \phi] \leq \min \left\{ \frac{c_{\mathcal{V}}(t, u)\phi(\mathcal{U}) + c_{\mathcal{V}}(t, u + 1)\phi(\mathcal{V})}{\phi(\mathcal{U}) + \phi(\mathcal{V})} \phi(\mathcal{V}) + \frac{c_{\mathcal{V}}(t, u)\phi(\mathcal{U}) + c_{\mathcal{U}}(t, u + 1)\phi(\mathcal{V})}{\phi(\mathcal{U}) + \phi(\mathcal{V})} \rho(\mathcal{U}), \phi(\mathcal{U}) + \phi(\mathcal{V}) \right\}.$$

Proof. We consider the two cases in which the first of the $t + 1$ new centers is chosen from either the covered set \mathcal{V} or the uncovered set \mathcal{U} , similar to the proof of Lemma 4. Denote by ϕ^1 the potential after adding the first new center.

Covered case: This case occurs with probability $\phi(\mathcal{V})/\phi$ and leaves the covered and uncovered sets unchanged. We then invoke Lemma 3 with $(t, u + 1)$ (one fewer center to add) and ϕ^1 playing the role of ϕ . The contribution to $\mathbb{E}[\phi' \mid \phi]$ from this case is then bounded by

$$\begin{aligned} & \frac{\phi(\mathcal{V})}{\phi} (c_{\mathcal{V}}(t, u + 1)\phi^1(\mathcal{V}) + c_{\mathcal{U}}(t, u + 1)\rho(\mathcal{U})) \\ & \leq \frac{\phi(\mathcal{V})}{\phi} (c_{\mathcal{V}}(t, u + 1)\phi(\mathcal{V}) + c_{\mathcal{U}}(t, u + 1)\rho(\mathcal{U})), \end{aligned} \quad (7)$$

noting that $\phi^1(\mathcal{S}) \leq \phi(\mathcal{S})$ for any set \mathcal{S} .

Uncovered case: We consider each uncovered cluster $\mathcal{A} \subseteq \mathcal{U}$ separately. With probability $\phi(\mathcal{A})/\phi$, the first new center is selected from \mathcal{A} , moving \mathcal{A} from the uncovered to the covered set and reducing the number of uncovered clusters by one. Applying Lemma 3 for (t, u) , the contribution to $\mathbb{E}[\phi' \mid \phi]$ is bounded by

$$\frac{\phi(\mathcal{A})}{\phi} [c_{\mathcal{V}}(t, u) (\phi^1(\mathcal{V}) + \phi^1(\mathcal{A})) + c_{\mathcal{U}}(t, u)(\rho(\mathcal{U}) - \rho(\mathcal{A}))].$$

Taking the expectation with respect to possible centers in \mathcal{A} and using Lemma 2 and $\phi^1(\mathcal{V}) \leq \phi(\mathcal{V})$, we obtain the further bound

$$\frac{\phi(\mathcal{A})}{\phi} [c_{\mathcal{V}}(t, u)(\phi(\mathcal{V}) + \rho(\mathcal{A})) + c_{\mathcal{U}}(t, u)(\rho(\mathcal{U}) - \rho(\mathcal{A}))].$$

Summing over $\mathcal{A} \subseteq \mathcal{U}$ yields

$$\begin{aligned} & \frac{\phi(\mathcal{U})}{\phi} (c_{\mathcal{V}}(t, u)\phi(\mathcal{V}) + c_{\mathcal{U}}(t, u)\rho(\mathcal{U})) + \frac{c_{\mathcal{V}}(t, u) - c_{\mathcal{U}}(t, u)}{\phi} \sum_{\mathcal{A} \subseteq \mathcal{U}} \phi(\mathcal{A})\rho(\mathcal{A}) \\ & \leq \frac{\phi(\mathcal{U})}{\phi} c_{\mathcal{V}}(t, u)(\phi(\mathcal{V}) + \rho(\mathcal{U})), \end{aligned} \quad (8)$$

using the inner product bound (18).

The result follows from summing (7) and (8) and combining with the trivial bound $\mathbb{E}[\phi' \mid \phi] \leq \phi = \phi(\mathcal{U}) + \phi(\mathcal{V})$. \square

As noted above, the bound in Lemma 5 depends on $\phi(\mathcal{U})$, the potential over uncovered clusters. This quantity can be arbitrarily large or small. In the next lemma, $\phi(\mathcal{U})$ is eliminated by maximizing with respect to it.

Lemma 6. *Assume that Lemma 3 holds for (t, u) and $(t, u + 1)$ with $c_{\mathcal{V}}(t, u + 1) \geq 1$. Then for the case $(t + 1, u + 1)$ in the sense of Lemma 5,*

$$\mathbb{E}[\phi' \mid \phi] \leq \frac{1}{2} c_{\mathcal{V}}(t, u)(\phi(\mathcal{V}) + \rho(\mathcal{U})) + \frac{1}{2} \max \left\{ c_{\mathcal{V}}(t, u)(\phi(\mathcal{V}) + \rho(\mathcal{U})), \sqrt{Q} \right\},$$

where

$$\begin{aligned} Q = & (c_{\mathcal{V}}(t, u)^2 - 4c_{\mathcal{V}}(t, u) + 4c_{\mathcal{V}}(t, u + 1)) \phi(\mathcal{V})^2 \\ & + 2(c_{\mathcal{V}}(t, u)^2 - 2c_{\mathcal{V}}(t, u) + 2c_{\mathcal{U}}(t, u + 1)) \phi(\mathcal{V})\rho(\mathcal{U}) + c_{\mathcal{V}}(t, u)^2 \rho(\mathcal{U})^2. \end{aligned}$$

Proof. The result is obtained by maximizing the bound in Lemma 5 with respect to $\phi(\mathcal{U})$. Let $B_1(\phi(\mathcal{U}))$ and $B_2(\phi(\mathcal{U}))$ denote the two terms in the minimum. The derivative of $B_1(\phi(\mathcal{U}))$ is given by

$$B_1'(\phi(\mathcal{U})) = \frac{\phi(\mathcal{V})}{(\phi(\mathcal{U}) + \phi(\mathcal{V}))^2} [(c_{\mathcal{V}}(t, u) - c_{\mathcal{V}}(t, u + 1))\phi(\mathcal{V}) + (c_{\mathcal{V}}(t, u) - c_{\mathcal{U}}(t, u + 1))\rho(\mathcal{U})],$$

which does not change sign as a function of $\phi(\mathcal{U})$. The two cases $B_1'(\phi(\mathcal{U})) \geq 0$ and $B_1'(\phi(\mathcal{U})) < 0$ are considered separately below. Taking the maximum of the resulting bounds (9), (10) establishes the lemma.

Case $B_1'(\phi(\mathcal{U})) \geq 0$: Both $B_1(\phi(\mathcal{U}))$ and $B_2(\phi(\mathcal{U}))$ are non-decreasing functions of $\phi(\mathcal{U})$. The former has the finite supremum

$$c_{\mathcal{V}}(t, u)(\phi(\mathcal{V}) + \rho(\mathcal{U})), \quad (9)$$

whereas the latter increases without bound. Therefore $B_1(\phi(\mathcal{U}))$ eventually becomes the smaller of the two and (9) can be taken as an upper bound on $\min\{B_1(\phi(\mathcal{U})), B_2(\phi(\mathcal{U}))\}$.

Case $B_1'(\phi(\mathcal{U})) < 0$: At $\phi(\mathcal{U}) = 0$, we have $B_1(0) = c_{\mathcal{V}}(t, u + 1)\phi(\mathcal{V}) + c_{\mathcal{U}}(t, u + 1)\rho(\mathcal{U})$ and $B_2(0) = \phi(\mathcal{V})$. The assumption $c_{\mathcal{V}}(t, u + 1) \geq 1$ implies that $B_1(0) \geq B_2(0)$. Since $B_1(\phi(\mathcal{U}))$ is now a decreasing function, the two functions must intersect and the point of intersection then provides an upper bound on $\min\{B_1(\phi(\mathcal{U})), B_2(\phi(\mathcal{U}))\}$.

Solving for the intersection leads after some algebra to a quadratic equation in $\phi(\mathcal{U})$:

$$0 = \phi(\mathcal{U})^2 + [2\phi(\mathcal{V}) - c_{\mathcal{V}}(t, u)(\phi(\mathcal{V}) + \rho(\mathcal{U}))]\phi(\mathcal{U}) + \phi(\mathcal{V})(\phi(\mathcal{V}) - c_{\mathcal{V}}(t, u + 1)\phi(\mathcal{V}) - c_{\mathcal{U}}(t, u + 1)\rho(\mathcal{U})).$$

Again by the assumption $c_{\mathcal{V}}(t, u + 1) \geq 1$, the constant term in this quadratic equation is non-positive, implying that one of the roots is also non-positive and can be discarded. The remaining positive root is given by

$$\phi(\mathcal{U}) = \frac{1}{2}c_{\mathcal{V}}(t, u)(\phi(\mathcal{V}) + \rho(\mathcal{U})) - \phi(\mathcal{V}) + \frac{1}{2}\sqrt{Q}$$

after simplifying the discriminant to match the stated expression for Q . Evaluating either $B_1(\phi(\mathcal{U}))$ or $B_2(\phi(\mathcal{U}))$ at this root gives

$$\frac{1}{2}c_{\mathcal{V}}(t, u)(\phi(\mathcal{V}) + \rho(\mathcal{U})) + \frac{1}{2}\sqrt{Q}. \quad (10)$$

□

The bound in Lemma 6 is a function of $\phi(\mathcal{V})$ and $\rho(\mathcal{U})$ only but is nonlinear, in contrast to the desired form in Lemma 3. The next step is to linearize the bound by imposing additional conditions (5) on the coefficients.

Lemma 7. *Assume that Lemma 3 holds for (t, u) and $(t, u + 1)$ with coefficients satisfying (5). Then for the case $(t + 1, u + 1)$ in the sense of Lemma 5,*

$$\mathbb{E}[\phi' \mid \phi] \leq \frac{1}{2} \left[c_{\mathcal{V}}(t, u) + (c_{\mathcal{V}}(t, u)^2 + 4 \max\{c_{\mathcal{V}}(t, u + 1) - c_{\mathcal{V}}(t, u), 0\})^{1/2} \right] \phi(\mathcal{V}) + c_{\mathcal{V}}(t, u)\rho(\mathcal{U}).$$

Proof. It suffices to linearize the \sqrt{Q} term in Lemma 6. In particular, we aim to bound the quadratic function Q from above by the square $(a\phi(\mathcal{V}) + b\rho(\mathcal{U}))^2$ for all $\phi(\mathcal{V}), \rho(\mathcal{U})$ and some choice of $a, b \geq 0$. The cases $\phi(\mathcal{V}) = 0$ and $\rho(\mathcal{U}) = 0$ require that

$$\begin{aligned} a^2 &\geq c_{\mathcal{V}}(t, u)^2 + 4(c_{\mathcal{V}}(t, u + 1) - c_{\mathcal{V}}(t, u)), \\ b^2 &\geq c_{\mathcal{V}}(t, u)^2. \end{aligned}$$

Setting these inequalities to equalities, the remaining condition for the cross-term is

$$ab \geq c_{\mathcal{V}}(t, u)^2 + 2(c_{\mathcal{U}}(t, u + 1) - c_{\mathcal{V}}(t, u)).$$

Equivalently for $a, b \geq 0$,

$$\begin{aligned} a^2 b^2 &= (c_{\mathcal{V}}(t, u)^2 + 4(c_{\mathcal{V}}(t, u + 1) - c_{\mathcal{V}}(t, u))) c_{\mathcal{V}}(t, u)^2 \\ &\geq (c_{\mathcal{V}}(t, u)^2 + 2(c_{\mathcal{U}}(t, u + 1) - c_{\mathcal{V}}(t, u)))^2. \end{aligned}$$

We rearrange to obtain

$$\begin{aligned} 4(c_{\mathcal{V}}(t, u + 1) - c_{\mathcal{V}}(t, u))c_{\mathcal{V}}(t, u)^2 &\geq 4c_{\mathcal{V}}(t, u)^2(c_{\mathcal{U}}(t, u + 1) - c_{\mathcal{V}}(t, u)) + 4(c_{\mathcal{U}}(t, u + 1) - c_{\mathcal{V}}(t, u))^2, \\ (c_{\mathcal{V}}(t, u + 1) - c_{\mathcal{U}}(t, u + 1))c_{\mathcal{V}}(t, u)^2 &\geq (c_{\mathcal{U}}(t, u + 1) - c_{\mathcal{V}}(t, u))^2, \end{aligned}$$

the last of which is true by assumption (5). Thus we conclude that

$$\sqrt{Q} \leq \sqrt{c_{\mathcal{V}}(t, u)^2 + 4(c_{\mathcal{V}}(t, u + 1) - c_{\mathcal{V}}(t, u))\phi(\mathcal{V}) + c_{\mathcal{V}}(t, u)\rho(\mathcal{U})}.$$

Combining this last inequality with Lemma 6 proves the result. \square

Given conditions (5) and Lemma 7, the inductive step for Lemma 3 can be completed by defining $c_{\mathcal{V}}(t + 1, u + 1)$ and $c_{\mathcal{U}}(t + 1, u + 1)$ recursively as in (6). \square

Equations (5) and (6) provide sufficient conditions on the coefficients $c_{\mathcal{V}}(t, u)$ and $c_{\mathcal{U}}(t, u)$ to establish Lemma 3 by induction. Section 4.1.3 shows that these conditions are satisfied by (2). To motivate the functional form chosen in (2), we first explore the behavior of solutions that satisfy (6) in particular. This is done by treating (6) as a recursion, taking the inequalities to be equalities, and numerically evaluating $c_{\mathcal{V}}(t + 1, u + 1)$ and $c_{\mathcal{U}}(t + 1, u + 1)$ starting from the base cases (3) and (4) as boundary conditions. More specifically, the computation is carried out as an outer loop over increasing u starting from $u + 1 = 1$, and an inner loop over t starting from $t = u + 1$. Figure 1 plots the resulting values for $c_{\mathcal{V}}(t, u)$ over the region $t \geq u$ ($c_{\mathcal{U}}(t, u)$ is simply a shifted copy). The most striking feature of Figure 1 is that the level contours appear to be lines $t \propto u$ emanating from the origin. Sampling values at multiple points (t, u) suggests that $c_{\mathcal{V}}(t, u) \approx t/(t - u)$. The plot also has the properties that $c_{\mathcal{V}}(t, u)$ is decreasing in t for fixed u and increasing in u for fixed t . These observations lead to the functional form for $c_{\mathcal{V}}(t, u)$ proposed in Section 4.1.3.

As for conditions (5), it can be verified directly using the base cases (3) and (4) that they are satisfied for $(t, u) = (1, 0)$. The subsequent numerical values in Figure 1 were found to satisfy (5) for all $t > u$ as well. This suggests that recursion (6) is self-perpetuating in the sense that if (5) are satisfied for (t, u) , then the values for $c_{\mathcal{V}}(t + 1, u + 1)$, $c_{\mathcal{U}}(t + 1, u + 1)$ resulting from (6) will also satisfy (5) for $(t + 1, u)$ and $(t + 1, u + 1)$, i.e. points to the right and upper-right. This self-perpetuating property is not proven however in the present paper. Instead, it is shown that the proposed functional form (11) satisfies (5) directly.

4.1.3 Proof with specific form for coefficients

We now prove that Lemma 3 holds for coefficients $c_{\mathcal{V}}(t, u)$, $c_{\mathcal{U}}(t, u)$ given by (11) below. These expressions are more general than (2) and are based on the observations drawn from Figure 1.

$$c_{\mathcal{V}}(t, u) = \frac{t + au + b}{t - u + b} = 1 + \frac{(a + 1)u}{t - u + b}, \quad t \geq u, \quad (11a)$$

$$c_{\mathcal{U}}(t, u) = \begin{cases} c_{\mathcal{V}}(t - 1, u - 1), & t \geq u > 0, \\ 0, & t \geq u = 0. \end{cases} \quad (11b)$$

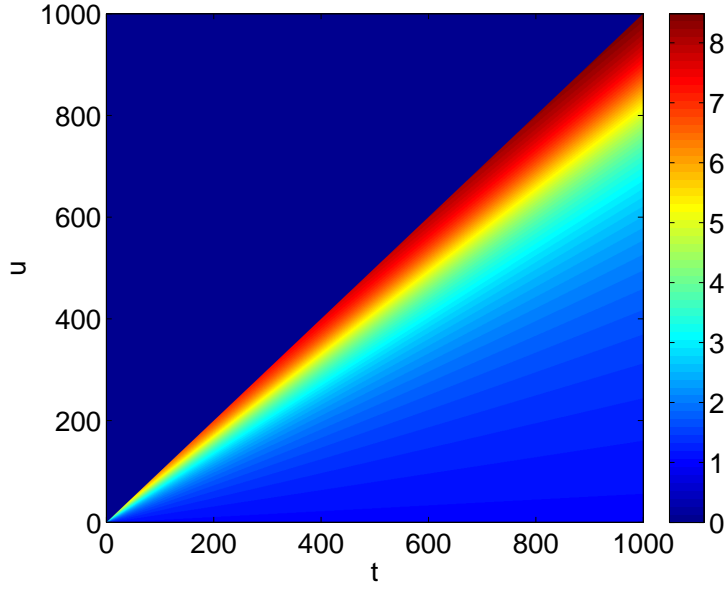


Figure 1: Coefficient $c_V(t, u)$ evaluated numerically in the region $t \geq u$ using recursion (6a) (treated as an equality) with boundary conditions (3) and (4). The numerical values approximate the function $t/(t - u)$.

Here a and b are parameters introduced to add flexibility to the basic form $t/(t - u)$ suggested by Figure 1, subject to the constraints $a > -1$, $b > 0$,

$$a + 1 \geq b, \quad (12a)$$

$$ab \geq 1. \quad (12b)$$

Equation (2) is obtained at the end from (11) by optimizing the parameters a and b . Note that with $a + 1 > 0$, (11a) is decreasing in t for fixed $u > 0$ and increasing in u for fixed t .

Given the inductive approach and the results established in Sections 4.1.1 and 4.1.2, the proof requires the remaining steps below. First, it is shown that the base cases (3), (4) from Section 4.1.1 imply that Lemma 3 is true for the same base cases but with $c_V(t, u)$, $c_U(t, u)$ given by (11) instead. Second, (11) is shown to satisfy conditions (5) for all $t > u$, thus permitting Lemma 7 to be used. Third, (11) is also shown to satisfy (6), which combined with Lemma 7 completes the induction.

Considering the base cases, for $u = 0$, (3) and (11) coincide so there is nothing to prove. For the case $t = u$, $u \geq 1$, Lemma 3 with coefficients given by (4) implies the same with coefficients given by (11) provided that

$$(1 + H_u)\phi(\mathcal{V}) + (1 + H_{u-1})\rho(\mathcal{U}) \leq \left(1 + \frac{(a+1)u}{b}\right)\phi(\mathcal{V}) + \left(1 + \frac{(a+1)(u-1)}{b}\right)\rho(\mathcal{U})$$

$$\quad \quad \quad \forall \phi(\mathcal{V}), \rho(\mathcal{U}).$$

This in turn is ensured if the coefficients satisfy $H_u \leq (a+1)u/b$ for all $u \geq 1$. The most stringent case is $u = 1$ and is met by assumption (12a).

For the second step of establishing (5), it is clear that (5a) is satisfied by (11a). A direct calculation presented in Appendix B shows that (5b) is also true.

Lemma 8. Condition (5b) is satisfied for all $t > u$ if $c_V(t, u)$ and $c_U(t, u)$ are given by (11) and (12b) holds.

Similarly for the third step, it suffices to show that (11a) satisfies recursion (6a) since (11b) automatically satisfies (6b). A proof is provided in Appendix C.

Lemma 9. Recursion (6a) is satisfied for all $t > u$ if $c_V(t, u)$ is given by (11a) and (12b) holds.

Having shown that Lemma 3 is true for coefficients given by (11) and (12), the specific expressions in (2) are obtained by minimizing $c_V(t, u)$ in (11a) with respect to a, b , subject to (12). For fixed a , minimizing with respect to b yields $b = a + 1$ in light of (12a), and

$$c_V(t, u) = 1 + \frac{(a+1)u}{t-u+(a+1)}.$$

Minimizing with respect to a then results in $a(a+1) = 1$ from (12b). The solution satisfying $a > -1$ is $a = \varphi - 1$ and $b = \varphi$. \square

4.2 Proof of Theorem 1

Denote by $n_{\mathcal{A}}$ the number of points in optimal cluster \mathcal{A} . In the first iteration of Algorithm 1, the first cluster center is selected from some \mathcal{A} with probability $n_{\mathcal{A}}/n$. Conditioned on this event, Lemma 3 is applied with covered set $\mathcal{V} = \mathcal{A}$, $u = k - 1$ uncovered clusters, and $t = \beta k - 1$ remaining cluster centers. This bounds the final potential ϕ' as

$$\mathbb{E}[\phi' \mid \phi] \leq c_V(\beta k - 1, k - 1)\phi(\mathcal{A}) + c_U(\beta k - 1, k - 1)(\rho - \rho(\mathcal{A}))$$

where $c_V(t, u)$, $c_U(t, u)$ are given by (2). Taking the expectation over possible centers in \mathcal{A} and using Lemma 1,

$$\mathbb{E}[\phi' \mid \mathcal{A}] \leq r_u^{(\ell)} c_V(\beta k - 1, k - 1)\phi^*(\mathcal{A}) + c_U(\beta k - 1, k - 1)(\rho - \rho(\mathcal{A})).$$

Taking the expectation over clusters \mathcal{A} and recalling that $\rho = r_D^{(\ell)}\phi^*$,

$$\mathbb{E}[\phi'] \leq r_D^{(\ell)} c_U(\beta k - 1, k - 1)\phi^* - C \sum_{\mathcal{A}} \frac{n_{\mathcal{A}}}{n} \phi^*(\mathcal{A}), \quad (13)$$

where

$$C = r_D^{(\ell)} c_U(\beta k - 1, k - 1) - r_u^{(\ell)} c_V(\beta k - 1, k - 1).$$

Next we aim to further bound the last term in (13). Using (2) and $r_D^{(\ell)} = 2^\ell r_u^{(\ell)}$ from Lemma 2,

$$\begin{aligned} C &= r_u^{(\ell)} \left(2^\ell c_U(\beta k - 1, k - 1) - c_V(\beta k - 1, k - 1) \right) \\ &= r_u^{(\ell)} \frac{2^\ell ((\beta - 1)k + \varphi(k - 1)) - (\beta - 1 + \varphi)k}{(\beta - 1)k + \varphi} \\ &= r_u^{(\ell)} \frac{(2^\ell - 1)(\beta - 1)k + \varphi((2^\ell - 1)(k - 1) - 1)}{(\beta - 1)k + \varphi}. \end{aligned}$$

The last expression for C is seen to be non-negative for $\beta \geq 1$, $k \geq 2$, and $\ell \geq 1$. Furthermore, since $n_{\mathcal{A}} = 1$ (a singleton cluster) implies that $\phi^*(\mathcal{A}) = 0$, we have

$$\sum_{\mathcal{A}} n_{\mathcal{A}} \phi^*(\mathcal{A}) = \sum_{\mathcal{A}: n_{\mathcal{A}} \geq 2} n_{\mathcal{A}} \phi^*(\mathcal{A}) \geq 2\phi^*, \quad (14)$$

with equality if ϕ^* is completely concentrated in clusters of size 2. Substituting (2b) and (14) into (13), we obtain

$$\frac{\mathbb{E}[\phi']}{\phi^*} \leq r_D^{(\ell)} \left(1 + \frac{\varphi(k-2)}{(\beta-1)k + \varphi} \right) - \frac{2C}{n}. \quad (15)$$

The last step is to recall Arthur and Vassilvitskii [2007, Theorems 3.1 and 5.1], which together state that

$$\frac{\mathbb{E}[\phi']}{\phi^*} \leq r_D^{(\ell)} (1 + H_{k-1}) \quad (16)$$

for ϕ' resulting from selecting exactly k cluster centers. In fact, (16) also holds for βk centers, $\beta \geq 1$, since adding centers cannot increase the potential. The proof is completed by taking the minimum of (15) and (16). \square

5 Conclusion and Future Work

This paper has shown that simple D^ℓ sampling algorithms, including k -means++, are guaranteed in expectation to attain a constant-factor bi-criteria approximation to an optimal clustering. The contributions herein extend and improve upon previous results concerning D^ℓ sampling [Arthur and Vassilvitskii, 2007, Aggarwal et al., 2009].

As noted in Section 3, the constant $r_D^{(\ell)}$ in Theorem 1 and Corollary 1 represents an opportunity to further improve the approximation bounds. One possibility is to tighten Lemmas 3.2 and 5.1 in Arthur and Vassilvitskii [2007], which are the lemmas responsible for the $r_D^{(\ell)}$ factor. A more significant improvement may result from considering not only the covering of optimal clusters by at least one cluster center, but also the effect of selecting more than one center from a single optimal cluster. As the number of selected centers increases, an approximation factor analogous to $r_D^{(\ell)}$ would be expected to decrease. Analysis of algorithms with similar simplicity to D^ℓ sampling is also of interest.

References

- A. Aggarwal, A. Deshpande, and R. Kannan. Adaptive sampling for k -means clustering. In *Proceedings of the 12th International Workshop and 13th International Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 15–28, August 2009.
- N. Ailon, R. Jaiswal, and C. Monteleoni. Streaming k -means approximation. In *Advances in Neural Information Processing Systems 22*, pages 10–18, December 2009.
- D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, May 2009.
- D. Arthur and S. Vassilvitskii. k -means++: The advantages of careful seeding. In *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, January 2007.
- V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k -median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, March 2004.
- P. Awasthi, M. Charikar, R. Krishnaswamy, and A. K. Sinop. The hardness of approximation of Euclidean k -means. In *Proceedings of the 31st International Symposium on Computational Geometry*, pages 754–767, June 2015.

- M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *Proceedings of the 34th ACM Symposium on Theory of Computing*, pages 250–257, May 2002.
- A. Bhattacharya, R. Jaiswal, and N. Ailon. A tight lower bound instance for k -means++ in constant dimension, volume 8402 of *Lecture Notes in Computer Science*, pages 7–22. Springer International, April 2014.
- T. Brunsch and H. Röglin. A bad instance for k -means++. *Theoretical Computer Science*, 505:19–26, September 2013.
- M. Charikar, S. Guha, E. Tardos, and D. B. Shmoys. A constant-factor approximation algorithm for the k -median problem. *Journal of Computer and System Sciences*, 65(1):129–149, August 2002.
- K. Chen. On coresets for k -median and k -means clustering in metric and Euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, September 2009.
- S. Dasgupta. The hardness of k -means clustering. Technical Report CS2008-0916, Department of Computer Science and Engineering, University of California, San Diego, 2008.
- W. F. de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani. Approximation schemes for clustering problems. In *Proceedings of the 35th ACM Symposium on Theory of Computing*, pages 50–58, June 2003.
- D. Feldman, M. Monemizadeh, and C. Sohler. A PTAS for k -means clustering based on weak coresets. In *Proceedings of the 23rd International Symposium on Computational Geometry*, pages 11–18, June 2007.
- S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O’Callaghan. Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):515–528, March 2003.
- S. Har-Peled and S. Mazumdar. On coresets for k -means and k -median clustering. In *Proceedings of the 36th ACM Symposium on Theory of Computing*, pages 291–300, June 2004.
- M. Inaba, N. Katoh, and H. Imai. Applications of weighted Voronoi diagrams and randomization to variance-based k -clustering. In *Proceedings of the 10th International Symposium on Computational Geometry*, pages 332–339, 1994.
- A. K. Jain. Data clustering: 50 years beyond k -means. *Pattern Recognition Letters*, 31(8):651–666, June 2010.
- K. Jain and V. V. Vazirani. Approximation algorithms for metric facility location and k -median problems using the primal-dual schema and Lagrangian relaxation. *Journal of the ACM*, 48(2):274–296, March 2001.
- K. Jain, M. Mahdian, and A. Saberi. A new greedy approach for facility location problems. In *Proceedings of the 34th ACM Symposium on Theory of Computing*, pages 731–740, May 2002.
- R. Jaiswal, A. Kumar, and S. Sen. A simple D^2 -sampling based PTAS for k -means and other clustering problems. *Algorithmica*, 70(1):22–46, September 2014.
- T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k -means clustering. *Computational Geometry*, 28(2–3):89–112, June 2004.
- A. Kumar, Y. Sabharwal, and S. Sen. Linear-time approximation schemes for clustering problems in any dimensions. *Journal of the ACM*, 57(2):5:1–5:32, January 2010.

- S. Lloyd. Least squares quantization in PCM. Technical report, Bell Laboratories, 1957.
- S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.
- M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k -means problem is NP-hard. In *Proceedings of the 3rd International Workshop on Algorithms and Computation*, pages 274–285, February 2009.
- K. Makarychev, Y. Makarychev, M. Sviridenko, and J. Ward. A bi-criteria approximation algorithm for k means. Technical Report arXiv:1507.04227, August 2015.
- J. Matoušek. On approximate geometric k -clustering. *Discrete & Computational Geometry*, 24(1):61–84, January 2000.
- R. R. Mettu and C. G. Plaxton. Optimal time bounds for approximate clustering. *Machine Learning*, 56(1–3):35–60, June 2004.
- R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k -means problem. *Journal of the ACM*, 59(6):28, December 2012.

A Proof of Lemma 4

The proof follows the inductive proof of Arthur and Vassilvitskii [2007, Lemma 3.3] with the notational changes $\mathcal{X}_u \rightarrow \mathcal{U}$, $\mathcal{X}_c \rightarrow \mathcal{V}$, and $8\phi_{\text{OPT}} \rightarrow \rho$. For brevity, only the differences are presented.

For the first base case $t = 0$, $u > 0$, Arthur and Vassilvitskii [2007] already show that the lemma holds with coefficients $1 = 1 + H_0$, $0 = 1 + H_{-1}$, and $1 = (u - 0)/u$. Similarly for the second base case $t = u = 1$, Arthur and Vassilvitskii [2007] show that $\mathbb{E}[\phi' \mid \phi] \leq 2\phi(\mathcal{V}) + \rho(\mathcal{U}) = (1 + H_1)\phi(\mathcal{V}) + (1 + H_0)\rho(\mathcal{U})$, as required for the stronger version here.

For the first “covered” case considered in the inductive step, the argument is the same and the upper bound on the contribution to $\mathbb{E}[\phi' \mid \phi]$ is changed to

$$\frac{\phi(\mathcal{V})}{\phi} \left[(1 + H_{t-1})\phi(\mathcal{V}) + (1 + H_{t-2})\rho(\mathcal{U}) + \frac{u - t + 1}{u}\phi(\mathcal{U}) \right]. \quad (17)$$

For the second “uncovered” case, the first displayed expression in the right-hand column of Arthur and Vassilvitskii [2007, page 1030] becomes (after applying the bound $\sum_{a \in \mathcal{A}} p_a \phi_a \leq \rho(\mathcal{A})$ from Lemma 2)

$$\frac{\phi(\mathcal{A})}{\phi} \left[(1 + H_{t-1})(\phi(\mathcal{V}) + \rho(\mathcal{A})) + (1 + H_{t-2})(\rho(\mathcal{U}) - \rho(\mathcal{A})) + \frac{u - t}{u - 1}(\phi(\mathcal{U}) - \phi(\mathcal{A})) \right].$$

Summing over all uncovered clusters $\mathcal{A} \subseteq \mathcal{U}$, the contribution to $\mathbb{E}[\phi' \mid \phi]$ is bounded from above by

$$\begin{aligned} & \frac{\phi(\mathcal{U})}{\phi} \left[(1 + H_{t-1})\phi(\mathcal{V}) + (1 + H_{t-2})\rho(\mathcal{U}) + \frac{u - t}{u - 1}\phi(\mathcal{U}) \right] \\ & + \frac{1}{\phi} \left[(H_{t-1} - H_{t-2}) \sum_{\mathcal{A} \subseteq \mathcal{U}} \phi(\mathcal{A})\rho(\mathcal{A}) - \frac{u - t}{u - 1} \sum_{\mathcal{A} \subseteq \mathcal{U}} \phi(\mathcal{A})^2 \right]. \end{aligned}$$

The inner product above can be bounded as

$$\sum_{\mathcal{A} \subseteq \mathcal{U}} \phi(\mathcal{A})\rho(\mathcal{A}) \leq \phi(\mathcal{U})\rho(\mathcal{U}), \quad (18)$$

with equality if both $\phi(\mathcal{U})$, $\rho(\mathcal{U})$ are completely concentrated in the same cluster \mathcal{A} . The sum of squares term can be bounded using the power-mean inequality as in Arthur and Vassilvitskii [2007]. Hence the contribution to $\mathbb{E}[\phi' \mid \phi]$ is further bounded by

$$\frac{\phi(\mathcal{U})}{\phi} \left[(1 + H_{t-1})\phi(\mathcal{V}) + (1 + H_{t-1})\rho(\mathcal{U}) + \frac{u-t}{u}\phi(\mathcal{U}) \right]. \quad (19)$$

Summing the bounds in (17), (19), we have

$$\mathbb{E}[\phi' \mid \phi] \leq (1 + H_{t-1})\phi(\mathcal{V}) + \left(1 + \frac{\phi(\mathcal{V})H_{t-2} + \phi(\mathcal{U})H_{t-1}}{\phi} \right) \rho(\mathcal{U}) + \frac{u-t}{u}\phi(\mathcal{U}) + \frac{\phi(\mathcal{V})}{\phi} \frac{\phi(\mathcal{U})}{u}.$$

Recalling that $\phi = \phi(\mathcal{V}) + \phi(\mathcal{U})$, the right-hand side is seen to be increasing in $\phi(\mathcal{U})$. Taking the worst case as $\phi(\mathcal{U}) \rightarrow \phi$ gives

$$\begin{aligned} \mathbb{E}[\phi' \mid \phi] &\leq \left(1 + H_{t-1} + \frac{1}{u} \right) \phi(\mathcal{V}) + (1 + H_{t-1})\rho(\mathcal{U}) + \frac{u-t}{u}\phi(\mathcal{U}) \\ &\leq (1 + H_t)\phi(\mathcal{V}) + (1 + H_{t-1})\rho(\mathcal{U}) + \frac{u-t}{u}\phi(\mathcal{U}) \end{aligned}$$

since $1/u \leq 1/t$. This completes the induction. \square

B Proof of Lemma 8

Substituting (11) into the left-most factor in (5b),

$$\begin{aligned} c_{\mathcal{V}}(t, u+1) - c_{\mathcal{U}}(t, u+1) &= c_{\mathcal{V}}(t, u+1) - c_{\mathcal{V}}(t-1, u) \\ &= \frac{(a+1)(u+1)}{t-u-1+b} - \frac{(a+1)u}{t-1-u+b} \\ &= \frac{a+1}{t-u-1+b}. \end{aligned}$$

Similarly on the right-hand side of (5b),

$$\begin{aligned} c_{\mathcal{U}}(t, u+1) - c_{\mathcal{V}}(t, u) &= c_{\mathcal{V}}(t-1, u) - c_{\mathcal{V}}(t, u) \\ &= \frac{(a+1)u}{t-1-u+b} - \frac{(a+1)u}{t-u+b} \\ &= \frac{(a+1)u}{(t-u+b)(t-u-1+b)}. \end{aligned}$$

Hence

$$\begin{aligned} &(c_{\mathcal{V}}(t, u+1) - c_{\mathcal{U}}(t, u+1))c_{\mathcal{V}}(t, u)^2 - (c_{\mathcal{U}}(t, u+1) - c_{\mathcal{V}}(t, u))^2 \\ &= \frac{a+1}{t-u-1+b} \left(1 + 2\frac{(a+1)u}{t-u+b} + \frac{(a+1)^2u^2}{(t-u+b)^2} \right) - \frac{(a+1)^2u^2}{(t-u+b)^2(t-u-1+b)^2} \\ &= \frac{a+1}{t-u-1+b} \left(1 + 2\frac{(a+1)u}{t-u+b} \right) + \frac{(a+1)^2u^2[(a+1)(t-u-1+b)-1]}{(t-u+b)^2(t-u-1+b)^2}. \end{aligned} \quad (20)$$

The first of the two summands in (20) is positive for $t > u \geq 0$. The second summand is also non-negative as long as $(a+1)(t-u-1+b) \geq 1$. The most stringent case occurs for $t = u+1$ and is implied by (12b). We conclude that (20) is positive, i.e. (5b) holds. \square

C Proof of Lemma 9

As noted earlier, (11a) has the property that $c_V(t, u+1) \geq c_V(t, u)$ for all t, u . Therefore (6a) is equivalent to

$$2c_V(t+1, u+1) - c_V(t, u) \geq \sqrt{c_V(t, u)^2 + 4(c_V(t, u+1) - c_V(t, u))}. \quad (21)$$

Substituting (11a) into the left-hand side,

$$\begin{aligned} 2c_V(t+1, u+1) - c_V(t, u) &= 1 + 2\frac{(a+1)(u+1)}{t-u+b} - \frac{(a+1)u}{t-u+b} \\ &= 1 + \frac{(a+1)(u+2)}{t-u+b}, \end{aligned}$$

which is seen to be positive for $t > u \geq 0$. Hence (21) is in turn equivalent to

$$(2c_V(t+1, u+1) - c_V(t, u))^2 \geq c_V(t, u)^2 + 4(c_V(t, u+1) - c_V(t, u)).$$

On the left-hand side,

$$(2c_V(t+1, u+1) - c_V(t, u))^2 = 1 + 2\frac{(a+1)(u+2)}{t-u+b} + \frac{(a+1)^2(u+2)^2}{(t-u+b)^2}. \quad (22)$$

On the right-hand side,

$$\begin{aligned} c_V(t, u+1) - c_V(t, u) &= \frac{(a+1)(u+1)}{t-u-1+b} - \frac{(a+1)u}{t-u+b} \\ &= \frac{(a+1)(t+b)}{(t-u+b)(t-u-1+b)} \\ &= \frac{a+1}{t-u+b} \left(1 + \frac{u+1}{t-u-1+b} \right), \\ c_V(t, u)^2 &= 1 + 2\frac{(a+1)u}{t-u+b} + \frac{(a+1)^2u^2}{(t-u+b)^2}, \end{aligned}$$

$$\begin{aligned} c_V(t, u)^2 + 4(c_V(t, u+1) - c_V(t, u)) \\ = 1 + 2\frac{(a+1)(u+2)}{t-u+b} + \frac{(a+1)^2u^2}{(t-u+b)^2} + 4\frac{(a+1)(u+1)}{(t-u+b)(t-u-1+b)}. \end{aligned} \quad (23)$$

Subtracting (23) from (22) yields

$$\begin{aligned} &\frac{4(a+1)^2(u+1)}{(t-u+b)^2} - 4\frac{(a+1)(u+1)}{(t-u+b)(t-u-1+b)} \\ &= 4\frac{(a+1)(u+1)[a(t-u-1+b)-1]}{(t-u+b)^2(t-u-1+b)}, \end{aligned}$$

which is non-negative provided that $a(t-u-1+b) \geq 1$. As in the proof of Lemma 8, the most stringent case occurs for $t = u+1$ and is covered by (12b). We conclude that (22) is at least as large as (23), i.e. (6a) holds. \square